LONG-RANGE FORECASTING
From Crystal Ball to Computer
Twelve

TESTING INPUTS

Contents

Validity and Reliability ...................... 323
Testing the Model Building Process ...................... 325
Testing the Variables ...................... 327
Testing the Relationships ...................... 329
Summary ...................... 331
Testing Inputs

The most essential qualification for a politician is the ability to foretell what will happen tomorrow, next month, and next year, and to explain afterwards why it did not happen.

Winston Churchill
(via Cetron and Ralph, 1983)

A famous forecast concluded that the national IQ would decline. The basic assumptions for this forecast were that people who had lower IQs had more children than those with higher IQs, and that IQ is largely hereditary. This forecast was supported by well-respected sociologists and other people who might be considered to be experts. What was wrong with this forecast? It did not forecast well. (However, by some measures of intelligence, national quotients have shown a decline since the early 1960s. RIMLAND [1981] reviews alternative explanations.) Still, the theory persisted, perhaps because the inputs were so appealing: everyone knows that those less intelligent multiply at a faster rate.

A careful examination of the inputs to this IQ forecast reveals that they are faulty. It is true that IQ is hereditary to a large extent. However, it is not true that less intelligent people multiply at a faster rate. (Actually, they can’t even add faster.) This false assumption arose because it was noted that larger families are produced by lower IQ people. However, the higher reproductive rate of those in the lower IQ groups who are parents is offset by the larger proportion of this group who never marry or who fail to reproduce at all. This explanation was first proposed by Willoughby and Cougan (1940) and later supported by Duncan (1952, 1969), Higgins, Reed, and Reed (1962), and OLNECK and WOLFE [1980].

The IQ forecasting model illustrates one of the uses of the examination of inputs, namely, to rule out models based on highly fallacious assumptions. Seldom, however, is a case so clear-cut that it will allow us to discard a model. In general, the testing of inputs is done to find ways to improve the model. Chapter 12 provides specific advice on testing inputs.

Forecasting models have two basic components: variables and relationships. Each component can be analyzed separately. In addition, you can examine the process that yields the variables and relationships. These various inputs are illustrated by the rectangles in Exhibit 12-1.

The concepts of validity and reliability are of particular importance
for analyzing inputs. The first part of the chapter describes these concepts. They are then applied to testing the process, the variables, and the relationships.

VALIDITY AND RELIABILITY

Techniques for improving the validity and reliability of forecasting methods were provided in Part II of this book. In this chapter, these concepts are discussed in more detail and their role in testing is described. This description draws upon Sellitz, Wrightsman, and Cook (1976).

Tests for validity may be classified into three categories: face validity, predictive validity, and construct validity. These tests are described along with the test of reliability.

The test of face validity is whether “people who should know” agree that something is reasonable. This is the weakest of the validity tests. However, it is one on which most of us place great reliance. (As a check on this, you might ask your friends about the national IQ forecasting model. Some people cling strongly to false but seemingly reasonable assumptions.) Kelly and Fiske (1950) suggest that a more appropriate name would be “faith validity.”

“People who should know” often do not know. Hotelling presents a case demonstrating the lack of a critical eye among academics. He was concerned that authors of textbooks did not understand the assumptions behind their materials. Consequently, they tended to introduce errors as the “facts” were copied from textbook to textbook. Hotelling wrote (1940, pp. 460–461):
One outstanding example is in certain formulae connected with the rank correlation coefficient, derived originally by Karl Pearson in 1907 and copied from textbook to textbook without adequate checking back. As one error after another was introduced in this process, the formulae presented to students . . . became less and less like Pearson's original equations.

To stem the flow of my own errors, I provide an Errata section in this Second Edition (LRF p. 450).

Mosier (1947) assessed the value of face validity by using two tests designed to select office workers. The two tests seemed to measure the same thing (i.e., the tests were the same, judged on face validity). You might examine them to see whether you think they measure the same skill:

### Alphabetizing Test

**Test 1.** Below are five names in random order. If the names were placed in strict alphabetical order, which name would be *third*?

1. John Meeder
2. James Medway
3. Thomas Madow
4. Catherine Meagan
5. Eleanor Meehand

**Test 2.** In the following items you have one name which is underlined and four other names in alphabetical order. If you were to put the underlined name into alphabetical series, indicate by the appropriate letter where it would go:

A. ______________________
   Richard Carreton
B. ______________________
   Roland Casstar
C. ______________________
   Jack Corson
D. ______________________
   Edward Cranston
E. ______________________

When these tests were used on the same 43 clerical workers, the correlation between them was .01. In other words, those who passed
test 1 were no more likely to pass test 2 than those who failed test 1. Also, the correlations between tests 1 and 2 and supervisors' ratings for 72 filing clerks were only .09 and .00, respectively.

Face validity suffers from the same problems as those associated with other judgmental data. As with judgmental forecasting, the process can be improved by adding structure. For example, a self-administered questionnaire asking experts to evaluate the validity of a set of assumptions is preferable to the use of unsolicited feedback. But even this is a weak test of validity. Strangely, few researchers go this far. The dominant approach to validity in many fields is the informal appeal to face validity.

Predictive validity, as applied to the inputs of the model, asks whether these inputs are valid for the forecast situation. For example, for a causal model, assessments would be made of the predictions of the causal variables. Also, to what extent are the causal relationships valid over the forecast horizon?

Construct validity asks whether a measurement measures what it claims to measure. Construct validity is assessed by using different methods to measure the construct and by examining the agreement among these measures. If these different approaches do, in fact, measure the same construct, they should agree with one another. This is another application of eclectic research.

Given that the inputs are valid, it is useful to ask whether they are reliable. Do repeated measures of the same input provide similar results? For example, it is useful to show that the procedure used to measure a given variable can be performed independently by different researchers and yield similar results each time. Reliability can also be assessed by having the same researchers repeat the procedure at different times.

The distinction between the tests for reliability and construct validity can be stated as follows: Reliability involves the agreement among similar ways of measuring an input; construct validity involves the agreement among dissimilar ways of measuring an input. This distinction is obviously one of degree, and often the differences are not clear-cut in practice. Still, the distinction, is useful in developing testing procedures. This should become clearer in the following sections where examples are provided.

TESTING THE MODEL BUILDING PROCESS

Obviously, many subjective decisions must be made in building a forecasting model. If the process can be replicated, this gives added as-
insurance that the model will be useful. If not, the failure suggests that more effort is needed to make the model building process more explicit and structured.

The test of reliability asks whether another researcher can follow the specific process and obtain the same results. This requires full disclosure of the process. If you have ever tried to provide full disclosure of a process, you know that it is expensive. If you have tried to replicate work done by others, you know that few people are able to provide full disclosure. Sometimes it is not the researcher's fault; it may be due to limitations of time or budget.

In addition to testing the reliability of the model building process, one could also test its construct validity. This is done to see whether researchers can follow the general process. In other words, rather than trying to follow the specific steps, a researcher would start from scratch and independently develop a model. This is an expensive test because all of the work must be repeated. If similar results are obtained, one gains confidence in the model. If different results are obtained, one gains clues about critical aspects of the process. This latter benefit comes about, however, only if each researcher provides full disclosure.

Below, I present a study drawn from econometric models. The study suggests that the econometric model building process is replicable and valid. Three groups of researchers independently reached similar conclusions in this situation:

Armstrong and Grohman (1972) developed an econometric model to predict air travel in the United States. After this model was completed, we learned that the U.S. Civil Aeronautics Board (CAB) had previously developed an econometric model (Saginor, 1967). The two models proved to be highly similar. The selection of causal variables was similar: both models used population, average fare per passenger mile, and real disposable income per capita, although the Armstrong–Grohman model added speed and safety variables. The same functional form was used: a multiplicative model that examined changes in key causal variables. The signs of the relationships were the same, and the magnitudes were similar: for example, the CAB estimated price elasticity to be −1.2, and our estimate was −1.2; the CAB estimated income elasticity to be +1.1, and we estimated +0.5. Finally, the accuracy of unconditional forecasts from the CAB model was similar to that from our model.
Our tardy literature review turned up another bonus: a study of the U.S. air travel market by Mize and Ulveling (1970). They did not seem to be aware of Saginor's study and were obviously unaware of our study, yet they also came up with a model using similar variables, functional form, and elasticities. Furthermore, they achieved an average accuracy of 4% in forecasts for 1964, 1965, and 1966; this was comparable to the error from our model of 6%. Both models did much better than extrapolations during this period.

Replication of the model building process also helps to control for cheating and mistakes.

TESTING THE VARIABLES

The reliability and validity of the dependent variable (the variable to be forecast) is of utmost importance [KEREN and NEWMAN, 1978]. Failure to do an adequate job here can jeopardize the rest of the enterprise. For example, ALEXANDER and WILKINS [1982] found that performance ratings have little validity; their conclusion raises questions about much of the prior research on predicting performance of job applicants. In addition, one should test the reliability and validity of the causal variables.

Application of the tests for validity and reliability is relatively straightforward. For reliability, you examine the agreement among similar measures of a variable (see Curtis and Jackson 1962, for an early but still useful description of this strategy). Another measure of reliability can be obtained from the variation among sample observations. To test for validity, you can use the agreement among dissimilar measures of the same variable (construct validity), or the ability to forecast the causal variables over the forecast horizon (forecast validity). You could also conduct a mail survey of experts to determine what they think to be valid measures (Sherbini, 1967, used this approach to test face validity in a study of international markets).

The following two studies illustrate the application of tests of reliability and construct validity. (Enough has already been said about face validity, and details on the predictive validity of inputs are not discussed here because the process is the same as that used in Chapter 13.) The example from Rabinowitz and Rosenbaum is interesting. They
followed much of my advice (even anticipated what I would say!) and, as noted, did a thorough job in assessing reliability. Still they were unable to predict student–teacher rapport in the classroom. In any event, this study illustrates the testing of reliability. Such testing is common in psychology, but uncommon in other fields such as economics. My study of the photographic market illustrates the testing of construct validity, although it also contains elements of reliability testing:

Rabinowitz and Rosenbaum (1958) tested 49 student teachers in an attempt to predict student–teacher rapport when these people would be teaching in grades 3 to 6 one year later. For the dependent variable, rapport, each teacher was observed by six raters, and each rater worked alone for two half-hour periods. The dependent variable showed high interrater reliability. The causal variables were based on tests whose reliability had been assessed by others.

In Armstrong (1968a), three measures were obtained for the price of cameras: a Kodak index based on the Kodak Instamatic 104 camera plus 10 packs of “type 126” black and white film; a Polaroid index based on the model 104 camera and 10 packs of “type 107” black and white film; and a Canon index based on the QL 25 camera. If each of these indices is a valid measure of the price of still cameras, they should be in agreement on which countries have high prices and which have low prices. Data from 26 countries were used to assess the agreement, and a measure of $r^2$ was computed for each pair of indices. There was a substantial amount of construct validity. Measures of $r^2$ were as follows:

<table>
<thead>
<tr>
<th></th>
<th>Polaroid</th>
<th>Canon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kodak</td>
<td>.90</td>
<td>.79</td>
</tr>
<tr>
<td>Polaroid</td>
<td>—</td>
<td>.67</td>
</tr>
</tbody>
</table>

In addition, the reliability of the data was assessed by repeating our survey to obtain the price data. The same questionnaires were sent to the same importers in each country after a 6-month interval. The results were identical in most cases.
The various methods of validity and reliability testing are complements—not substitutes—of one another. In other words, a battery of tests should be used, and the input should be expected to do well on all. Ferber illustrated how misleading it would be to use a single measure:

In Ferber (1965), replies were received from 411 bank customers on the sizes of their savings deposits. The average deposit in this sample was calculated, and reliability was assessed for this average by using the variability among respondents (i.e., the traditional measure of sample reliability). Bank records showed that the true value was outside the 95% confidence interval (the reported deposit was half of the actual deposit). The measure of reliability provided a poor assessment in this case.

**TESTING THE RELATIONSHIPS**

The testing of relationships is relevant only for econometric, segmentation, and bootstrapping models, because it is here that relationships are explicitly examined.

To test the reliability of estimates of relationships, one examines the agreement among similar approaches. The most commonly used approach for assessing what would happen if the measurement process were repeated is to use standard errors of regression coefficients. A small standard error implies high reliability.

Although the standard error provides a simple and inexpensive measure, its shortcomings should be noted: it does not assess validity, and it overstates reliability because of the fitting process in the *calibration sample*\(^6\). The second problem can often be handled by splitting the data into two subsamples. The analysis, with all of the adjustments that inevitably occur, can be done on one of these subsamples. The resulting model can then be used with the second subsample to calculate standard errors. The disadvantage of this *split-sample*\(^6\) technique is that one must often deal with small samples. Of course, it is possible to use the split-sample technique with small samples, and then to combine both samples to obtain an estimate after having assessed reliability.

When the split-sample technique is not available, it is still possible to account for chance relationships that show up when many variables are considered. Other people have noted this problem long ago (Binder 1964, for example), and they have been kind enough to provide tables...
so that you can avoid being misled by the traditional $t$-statistic. [See McIntyre et al. 1983.] They adjust for the fact that many variables were called, but few were chosen for the model.

The test of construct validity helps to assess whether the relationship can be generalized to new situations (e.g., new time periods, new geographical regions, or new subjects). It also helps to assess whether uncertainty is high or whether biases exist in the data, thus helping to identify any further research that would be useful.

To assess construct validity, the relationship is measured by approaches that are as different as possible. The approaches can differ in many ways: different measures for the variables in the forecasting model, different time periods, different types of data, different analytical methods, or different subjects.

My study of the photographic market is used to illustrate tests of reliability, construct validity, and predictive validity of relationships:

The measurement of relationships in the study of the international photographic market (Armstrong, 1968a) was described briefly in Chapter 8 of LRF. The estimate of income elasticity was of particular interest, so four estimates were obtained. For the subjective estimate, reliability was estimated by me. Reliability estimates from regression analyses on three types of data were obtained from the standard errors of the regression coefficients. Construct validity was estimated by calculating the standard deviation among the four different estimates of income elasticity. The results were as follows:

<table>
<thead>
<tr>
<th>Source of Estimate</th>
<th>Income Elasticity</th>
<th>Test</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective</td>
<td>1.3</td>
<td>Reliability</td>
<td>0.2</td>
</tr>
<tr>
<td>Household survey</td>
<td>1.5</td>
<td>Reliability</td>
<td>0.2</td>
</tr>
<tr>
<td>Cross section of countries</td>
<td>0.9</td>
<td>Reliability</td>
<td>0.1</td>
</tr>
<tr>
<td>Longitudinal over countries</td>
<td>1.6</td>
<td>Reliability</td>
<td>1.0</td>
</tr>
<tr>
<td>All of the above</td>
<td>1.3</td>
<td>Construct validity</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Because incomes were rising over time, an assessment was made of the predictive validity of the estimate of income elasticity. The cross-sectional data were split into two groups: 15 countries
of low income and 15 of high income. The estimated income elasticity for the high-income countries was 1.01 (standard error = 0.35), while that for the low-income countries was 1.03 (standard error = 0.11). These results supported the assumption that the estimate of income elasticity would be valid as income increased over the 10-year forecast horizon.

SUMMARY

Tests on the inputs of a model help to identify weak areas so that improvements can be made. A second goal of input testing is to select the most useful models. This is of less importance; generally, the best one can do is to eliminate models with highly unreasonable assumptions.

Input testing can be used on three aspects of the inputs; testing the model building process, testing variables, and testing relationships. For each of these aspects, one can examine reliability and validity. The suggestions made for these analyses are summarized in Exhibit 12-2.

<table>
<thead>
<tr>
<th>Exhibit 12-2 METHODS FOR TESTING INPUTS TO A MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
</tr>
<tr>
<td>Model building process</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Variables</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Relationships</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>